Diseño y construcción de un modelo descriptivo de datos para el manejo de productos farmacéuticos de la Droguería Puerto Boyacá S.A.S.

Jhoan Sebastián Marín Valencia & Mauricio Castaño Gómez

Noviembre 2019

Universidad Tecnológica de Pereira

Facultad de Ingenierías

Ingeniería de Sistemas y Computación

Diseño y construcción de un modelo descriptivo de datos para el manejo de productos farmacéuticos de la Droguería Puerto Boyacá S.A.S.

Tesis de grado para optar al título de

Ingeniero de Sistemas y Computación

Jhoan Sebastián Marín Valencia & Mauricio Castaño Gómez

Noviembre 2019

Profesor guía

MSc. Carlos Andrés López

Universidad Tecnológica de Pereira

Facultad de Ingenierías

Ingeniería de Sistemas y Computación

Agradecimientos

Agradecimientos principales a Dios.

A todos los familiares por ser pilares fundamentales en la formación de nuestras vidas.

A nuestro profesor guía Carlos por su asesoramiento y ayuda durante la ejecución del presente trabajo.

A toda la planta docente de la Universidad Tecnológica de Pereira por la contribución de sus conocimientos y experiencias que nos permitieron formarnos profesionalmente.

Al gerente comercial de FarmIPS Juan Carlos por todas sus explicaciones y aportes para el desarrollo del proyecto.

A la representante de la empresa Droguería Puerto Boyacá S.A.S. Carmen Rosa Verano por su disposición y ayuda durante el trabajo.

Jhoan Sebastián Marín Valencia

Mauricio Castaño Gómez

Dedicatoria

A mi madre, padre, hermanos y mi novia Alejandra por todo su apoyo incondicional.

Mauricio Castaño Gómez

Este trabajo va dedicado a mis padres y hermanas que siempre han sido un motor en todo este trayecto y diseño de este proyecto de grado.

Jhoan Sebastián Marín Valencia

Tabla de contenido

Lista de ilustraciones

Lista de tablas

Resumen

En el presente trabajo de grado se lleva a cabo inicialmente una descripción de la situación actual de las pymes en Colombia, destacando sus problemas y situación más común. Luego se describe como la toma de decisiones es un proceso crítico para las empresas y cómo ha mejorado este a lo largo del tiempo, pasando a través de diferentes prácticas como la analítica de datos. Posteriormente se explica de forma general como una organización al tener más madurez analítica puede llegar a información que es vital para la toma de decisiones.

Posteriormente se aborda la situación actual de la mipyme Droguería Puerto Boyacá S.A.S., la cual presenta una dificultad para poder conocer su estado actual en el manejo de productos farmacéuticos, lo que genera a largo plazo una desventaja competitiva. A este problema es el al que se da solución en esta tesis, en la que se analizan los diferentes conjuntos de datos involucrados en el proceso.

Se describe primero cada uno de los conjuntos con los datos mínimos requeridos para un posterior análisis. El Registro Individual de Prestaciones de Salud (RIPS), es el primer conjunto el cual establece un límite en la cantidad de medicamentos a entregar por paciente. Planilla de entrega de medicamentos es el archivo que maneja la droguería para ir registrando las entregas realizadas. Y por último, el conjunto de datos de medicamentos que ha comprado la droguería.

Cada uno de estos conjuntos son procesados pasando por una limpieza de datos y luego por un análisis exploratorio, con el objetivo de dejar definidos los conjuntos requeridos con los datos mínimos necesarios para los análisis posibles a realizar que revelen información de valor al negocio. Con esto se procede a migrar a la base de datos MongoDB, la cual por sus

características, beneficios, herramientas, y diversas posibilidades se convierte en un gran aliado para que las mipymes puedan mejorar su administración de datos, llegando además a un estado de madurez analítica descriptiva en la que puedan observar lo que ocurre actualmente en su negocio, todo esto gracias a la herramienta open source de MongoDB Compass, la cual ofrece todas estas ventajas.

Por último, en el trabajo se muestran los resultados obtenidos tanto del proceso de limpieza de datos, el análisis exploratorio de datos y los beneficios del uso de Compass. Se finaliza brindando las conclusiones y hechos a destacar de todo el proceso llevado a cabo con la Droguería Puerto Boyacá S.A.S.

Introducción

Con el paso del tiempo han surgido diversas empresas, organizaciones, e instituciones, todas muy variadas entre sí con diferentes modelos de negocio, objetivos y necesidades, sin embargo, y sin duda alguna todas se enfrentan a situaciones que ya sean un problema o no requieren tomas de decisiones críticas que pueden afectar bastante al negocio y su entorno. Es entonces de esta forma como este proceso se vuelve de vital importancia para asegurar el éxito de la empresa y el desarrollo de su entorno.

"La toma de decisiones organizacionales se convirtió en objeto de estudio de diversas disciplinas científicas durante el pasado siglo XX" (Rodríguez Cruz & Pinto Molina, 2010). Por tal razón, han ido apareciendo técnicas, métodos, herramientas y enfoques distintos que buscan ser de gran soporte y apoyo, facilitando esta práctica y permitiendo que las decisiones sean las mejores posibles.

Por otra parte, las empresas recolectan y almacenan muchos datos que son valiosos, ya que estos al ser tratados o procesados revelan información que se vuelve indispensable para la toma de decisiones certera, rápida y eficaz (Rodríguez Cruz & Pinto Molina, 2010). Además, en la actualidad gracias a los big data (Mayer-Schönberger & Cukier, 2013) se abren nuevas posibilidades que en conjunto a las técnicas de análisis de datos, ciencia de datos, machine learning y demás, las organizaciones obtienen los resultados más deseados del proceso en cuestión.

Ahora bien, para que las instituciones y negocios puedan llegar hasta este punto en el que las técnicas y campos mencionados puedan ser de verdadera utilidad y generar un gran valor, es

necesario que estas avancen y evolucionen en sus aspectos tecnológicos y de tratamiento de datos. En la analítica de datos se presentó una escala que sirve de base para las organizaciones, permitiendo que éstas conozcan sus posibilidades y determinen las acciones pertinentes para avanzar y mejorar.

La escala de madurez de los datos (Elliott, 2013) propuesta por Gartner muestra los diferentes niveles con el valor de la información que se puede obtener en este estadio. Este aspecto tecnológico es para el área de tecnologías de la información (TI) en una organización de especial importancia, la cual debe buscar promover el avance de la empresa hacia un mejor nivel con información cada vez más valiosa que ayuda en la toma de las mejores decisiones, generando así un valor que le permite a la empresa crecer, ser más competente, prestar un mejor servicio, y demás.

Muchas de las pequeñas y medianas empresas (pymes y mipymes para micro empresas) en Colombia actualmente no cuentan con una buena infraestructura para las tecnologías de la información (TI), lo que resulta en una mal administración de los datos que estas poseen impidiendo así, obtener información con verdadera utilidad y valor para la toma de decisiones en la alta gerencia. Entre todas las pymes en el país se encuentra además una importante cantidad de estas que no están aún ni en el primer nivel de la escala (el nivel descriptivo), limitando de esta manera la posibilidad de expansión y mejora del negocio.

Esta situación se presenta en la empresa Droguería Puerto Boyacá SAS, y frente a esto se plantea ayudar al negocio para que avance en la administración de sus datos farmacéuticos y así pueda pasar de la desorganización y desinformación al estado descriptivo del modelo de madurez de los datos dado por Gartner, en el cual se podrán identificar problemas y ventajas actuales, fomentando el crecimiento de la empresa y la mejora de su entorno. Para cumplir este objetivo se

debe conocer su situación actual en TI con el propósito de poder realizar el diseño y la construcción de un modelo descriptivo de datos, llevar estos a una base de datos NoSQL en conjunto con la implantación de una interfaz web que posibilite el fácil acceso, comprensión y visualización de la información que generan las consultas a la base de datos.

Planteamiento del problema

Situación problema

Muchas de las pequeñas y medianas empresas o pymes según la encuesta anual de Brother International Corporation (Visión PYMES 2019, 2019), se enfrentan actualmente a diversos retos y problemas tanto organizacionales como tecnológicos que abarcan diferentes aspectos de infraestructura, personal, seguridad de la información, competitividad, y demás, lo que resulta en preocupación por la supervivencia y ventaja competitiva de parte de las mismas. En Colombia se evidencia en la encuesta el alto interés por la inversión en tecnología; con la compra de equipos de cómputo, contratación, y utilización de servicios en la nube, lo que demuestra evidencias de la transformación digital (Grupo Bit, 2018) por la que ha venido pasando Colombia desde ya hace algunos años, en donde las empresas cada vez más han dado cuenta de cómo la tecnología y prácticas como las de análisis de negocios y big data pueden llegar a ayudar y soportar la toma de decisiones.

"Sergio Gutiérrez, presidente de *Infórmese*, empresa especializada en Big Data, Colombia representa un panorama ideal para el surgimiento y el crecimiento del Big data y la Analítica de negocios" (Grupo Bit, 2018), y es que cada año se evidencia como en las encuestas las empresas tienden a la selección por soluciones tecnológicas y la contratación de personal de este campo, además la formación y especialización en Big Data, Businesss Intelligence (BI), y Ciencia de Datos ha estado creciendo.

Sin embargo, como mencionó Sergio Gutiérrez aún estamos en una etapa inicial, donde la mayoría de empresas apenas están empezando a dar primeros pasos hacia la implementación de las líneas de análisis mencionadas. Si bien, se puede inferir dado el mercado actual y la

encuestas anuales como muchas pymes actualmente no cuentan con una buena infraestructura y personal de TI que proporcione al negocio la correcta y buena administración de los datos que recolectan, impidiendo la posibilidad de obtener información valiosa que se requiere para el diagnóstico y posterior solución de problemas, esto es lo que se conoce en la Escala de Madurez de los Datos como el estado o nivel Descriptivo en el que las empresas pueden conocer su situación actual, un nivel inicial al cual están apuntando llegar.

Ahora bien, la Droguería Puerto Boyacá SAS es una mipyme dedicada mayormente al comercio de productos farmacéuticos y medicinales al por menor, cosméticos y artículos de tocador en establecimientos especializados, la cual a la fecha lleva funcionando más de 10 años en el mercado, pero a pesar de este tiempo presenta aún dificultades y problemas para determinar su situación actual en cuanto a información relacionada a la cantidad de ventas, ingresos, productos actuales, y como lo expresa la representante del negocio Carmen Rosa Verano; el desconocimiento de saber cuánto realmente es la ganancia (o posible pérdida) de los contratos relacionados a la distribución de productos farmacéuticos.

Definición del problema

Incapacidad de la Droguería Puerto Boyacá SAS para determinar con precisión la situación actual de ganancia o pérdida frente al manejo de productos farmacéuticos.

Análisis del problema

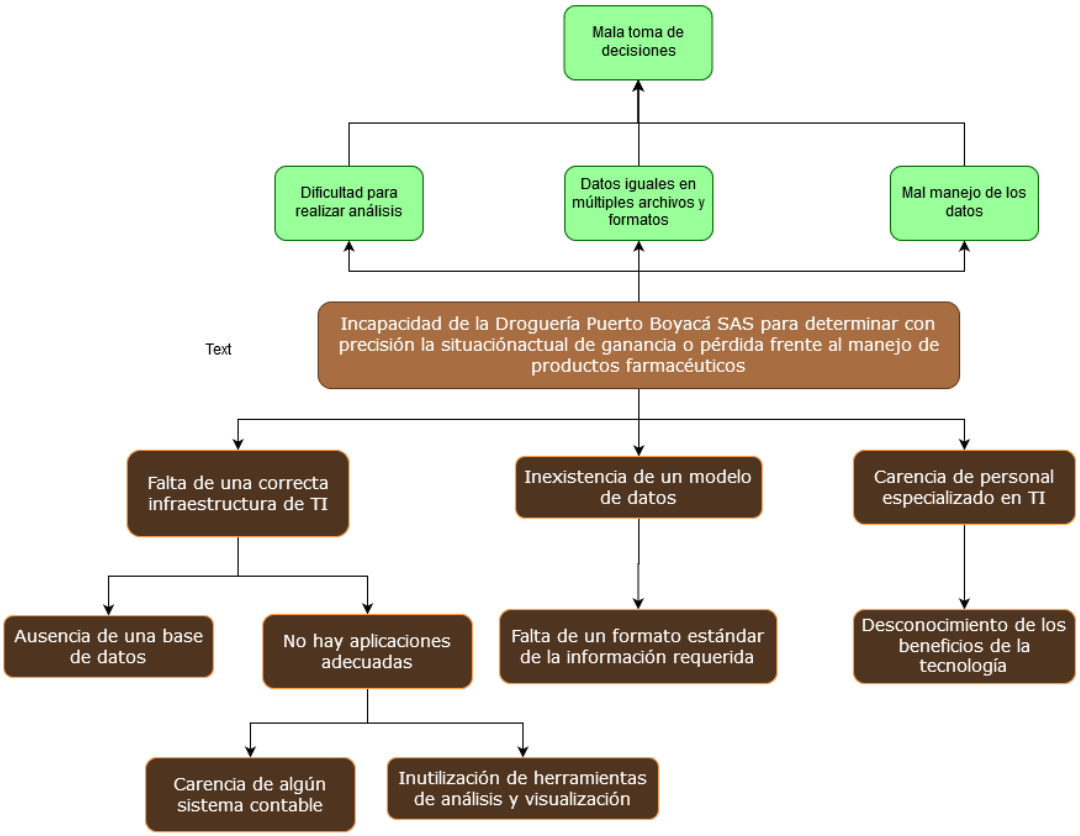

Ilustración 1. Árbol de problemas

Fuente: Elaboración propia

Actualmente la Droguería Puerto Boyacá SAS como negocio es incapaz de determinar y conocer verdaderamente cuáles son sus problemas o ventajas, frente al manejo de los productos farmacéuticos que comercializa. Esta situación ocurre por efecto de varias causas y problemas que la organización posee a nivel interno, estos son mostrados en la Ilustración 1.

Las tres fuentes principales de problemas son; la falta de una infraestructura, ausencia de un modelo de datos, y por último la carencia de personal especializado en TI o campos como la ciencia de datos, big data y afines. Cada una estos problemas son ya bastante conocidos y han sido solucionados por diversas organizaciones a través de variados enfoques o metodologías, como la implementación de las buenas prácticas de ITIL que fortalecen al negocio para hacer

frente a problemas actuales en la industria (Axelos, 2019), o la adopción de metodologías, estándares, modelos de madurez y guías (Beth Chrissis, Konrad, & Shrum, 2009) que buscan ayudar a la organización en la mejora de sus procesos y productos, en conjunto con la utilización de algún modelo de madurez analítica de datos que permita dar cuenta de su nivel, permitiendo emprender acciones para que la organización avance.

El principal problema a abordar en este trabajo es la creación del modelo de datos para los productos farmacéuticos y medicinales, determinando además lineamientos y acciones que le permitan a la Droguería Puerto Boyacá avanzar a un nivel analítico descriptivo. Sin embargo, para mantenerse en este punto es necesario dar solución permanente a los otros dos problemas, a los cuales se plantea brindar un primer apoyo proponiendo meramente una infraestructura básica para el modelo descriptivo de datos, así como la capacitación en la importancia de este.

Específicamente los problemas de infraestructura y carencia de personal requieren otro tipo de estudios, detalle e inversión que deben de ser evaluados por el negocio. Sin darle solución a estos el modelo a desarrollar puede ayudar, aunque aún existirían problemas para la realización de análisis, lo que resulta en últimas en las consecuencias mostradas en la ilustración 1. Si bien el modelo de datos ayuda a la realización de análisis descriptivos es necesario además que se mantengan los insumos de infraestructura (como servidores y aplicaciones) y personal.

Objetivos

Objetivo general

Diseñar y construir un modelo de datos descriptivo para los productos farmacéuticos administrados por la Droguería Puerto Boyacá SAS que le permita crecer en TI y avanzar al primer nivel en la escala de madurez de los datos.

Objetivos específicos

1. Reunir los datos y conocer el modelo de negocio de la Droguería Puerto Boyacá SAS.
2. Determinar la madurez de los datos de la Droguería Puerto Boyacá SAS.
3. Implementar una base de datos NoSQL.
4. Realizar un modelo de datos descriptivo para una base de datos NoSQL.

Justificación

En Colombia el 90% de las empresas son pymes o mipymes (COLOMBIA FINTECH, 2019), las cuales representan la mayor tasa de empleabilidad en el país (más del 65%) y aproximadamente el 30% del PIB nacional, siendo actualmente 2.540.953 mipymes las que figuran estas cifras. Esto genera especial importancia por tratar de que estas se mantengan y crezcan con objetivos finales de que mejore todo el territorio colombiano.

A pesar de esta relevancia, estudios actuales revelan que la mitad de las pymes se quiebran durante el primer año y sólo el 20% sobreviven al tercer año (COLOMBIA FINTECH, 2019), lo que presta preocupación por buscar soluciones que permitan mejorar esta situación. Si bien la mayor problemática presentada es la falta de financiación también se encuentran problemas relacionados a la toma de decisiones y a la inclusión de tecnología que hagan más eficiente el negocio.

Se ha de destacar que para el problema de financiación con el paso del tiempo se han abierto fuentes y fondos que buscan solventar esta necesidad , y por otra parte se evidencia actualmente el interés de las empresas por la inversión e inclusión de la tecnología (Visión PYMES 2019, 2019) en el negocio. Sin embargo, para los temas relacionados a la mejor toma de decisiones estratégicas aún se encuentran faltantes y malas prácticas en las que realizan este proceso sin basarse en datos certeros y precisos, ya que la mayoría de estas se encuentran aún fuera de una madurez analítica que les posibilite tomar las mejores decisiones con información que es clave para la ventaja competitiva (Elliott, 2013).

Y es que en la actualidad la información se convierte en un recurso esencial para una acertada, oportuna y rápida toma de decisiones estratégicas (Rodríguez Cruz & Pinto Molina, 2010), por tal razón la madurez analítica debe de empezar a incluirse en las pymes ya que con esto obtendrán información más valiosa que les permita superar la barrera que al día de hoy ha hecho que muchas de estas se pierdan.

Ahora bien, propender porque todas estas empresas mejoren en madurez analítica y tengan éxito es díficil debido a que se presentan bastantes modelos de negocio, y aún más contextos muy variados que hacen compleja una implementación generalizada. Ya bien explicó Garner que un primer reto necesario para la adopción de una infraestructura analítica es comprender todo el contexto de la organización (Elliott, 2013). Para esto Gartner además recomienda a las organizaciones la adopción de un framework con determinadas capacidades informaticas, el cual soporta varias categorías o casos de uso para las empresas.

Sin embargo, se ha de comprender que los diversos contextos de las pymes y mipymes colombianas pueden no acogerse todos al framework general propuesto por Gartner u otros más recientes. Son necesarios pasos y consideraciones adicionales a las necesidades específicas de cada empresa o grupos de estas bajo el mismo modelo de negocio, para determinar una metodología en conjunto de un modelo que facilite a todas las pymes incluirse en un mejor nivl analítico.

La creación del modelo descriptivo de datos para la Droguería Puerto Boyacá S.A.S pretende ser entonces un aporte particular al problema general abordado que presentan la mayoría de pymes en el país, permitiendo por ejemplo que otras empresas similares en contexto y con igual modelo de negocio tengan un referente con el cual determinar acciones o estrategias

con las cuales lleguen también al nivel descriptivo en la madurez analítica, abriéndose a estas todos los demás beneficios particulares que obtienen en dicho nivel.

De forma particular la droguería podrá entonces mediante el valor de la información que consigue en este nivel determinar mejores tomas de decisiones, que le ayuden ya sea a reducir pérdidas, mejorar procesos internos, obtener más ganancias, y hasta si lo desean continuar avanzando en tecnologías de la información, madurez analítica, inteligencia de negocios y demás campos afines. Aún más, con la mejora del negocio la empresa puede crecer, generando de esta forma más empleo, ayudando tributariamente y promoviendo tanto la mejora del sector en el que se encuentra como del territorio nacional, y por otra parte la atención a los usuarios puede mejorar llegando a contribuir así al cuidado de la salud de los ciudadanos.

Estado del arte

La madurez de los datos ha sido uno de los temas principales para el uso de herramientas de big data que permiten hacer análisis a grandes volúmenes de datos para que estos se conviertan en información de valor que sirva en la toma de decisiones o en la descripción de la situación actual de la organización.

Existen varios tipos de madurez de los datos tales como se puede observar en la opinión de Gama (Gamma, 2012) en la que expone la madurez de la calidad de los datos con el fin de apoyar los procesos internos de las organizaciones de forma que facilite identificar y fortalecer sus prácticas empresariales y la tecnología utilizada para el tratamiento de los datos, a fin de asegurar la calidad de los sistemas de información como un factor para la competitividad. Los modelos de madurez son de gran importancia debido a la aplicación en las organizaciones como un mecanismo que garantice la correcta incorporación de prácticas de gestión para llegar a un estado ideal que les permita a estas obtener una mayor competitividad.

Respecto a un modelo de calidad de los datos se han establecido diferentes propuestas alrededor del tema donde en su mayoría proceden al principio de caracterización de factores y diseño de herramientas para asegurar un nivel de aceptabilidad. Tal como se especifica en el paper de Gama a lo largo de la historia se han propuesto modelos para la medición de calidad que ayudan a la madurez de los datos: tal como lo es el planteamiento de un modelo de valoración de calidad de los datos, considerando una evaluación objetivo y una evaluación contextual, entre otros.

Center SinerTic Andino propone un modelo de medidas de información de calidad o por sus siglas en inglés Information Quality Metrics (IQM), la cual consiste en soluciones para el mejoramiento de la calidad de los datos que soportan los procesos estratégicos del negocio.

Teniendo en cuenta las métricas de calidad de información (IQM) como un modelo de madurez en calidad, existen ciertos procesos conocidos como “procesos de minería” los cuales representan los datos finales de mayor importancia para un método de análisis. Estos procesos de minería descubren datos de evento y procesos adyacentes que se involucran o intervienen en los resultados de un modelo. Muestra qué ha sido realizado durante los procesos y detectan discrepancias en lo que debería haber sido realizado. Tal como menciona Laura de Haan en su tesis de maestría “The integration of Big Data in purchasing, as designed in a new Big Data Purchasing Maturity model” (Haan, 2018) donde expone la importancia de procesos de minería de datos en los avances y ventajas competitivas que estos pueden generar dentro de una organización o entidad trabajando de la mano con el manejo o gestión de inventarios y con la cadena de suministros. Se evidencia que para el diseño de un modelo que defina la madurez de los datos es imprescindible usar procesos de minería o algoritmos de machine learning tales como arboles de clasificación, regresión logística, redes neuronales, clustering, con el fin de definir unos parámetros o establecer unos indicios que sean capaz de medir los aspectos importantes en la organización como los costos y los retos que cada una de estas presentan con base en el volumen de datos que posean. Teniendo claro esto, se permite un mejor diseño en un modelo de madurez de los datos según Haan.

En el 2013 se hace conocimiento del modelo de Gartner para la madurez analítico de los datos (Elliott, 2013) en donde se expresan 3 dimensiones como rango temporal que expresa el valor de los datos a través del tiempo, nivel de intervención humana la cual representa que tanto

debe interactuar con los datos para obtener información de relevancia y complejidad de los análisis matemáticos requeridos para sacar y obtener información de valor que responda las preguntas elocuentes a la organización. Este modelo presenta 4 niveles los cuales responden a diferentes preguntas de la organización (Omedes, 2017). Nivel Descriptivo, normalmente contesta a la pregunta "¿Qué ha ocurrido?", representa la información pasada mostrándose en análisis estadísticos y representa una intervención humana alta. Nivel Diagnóstico, busca responder a la pregunta "¿Por qué ha ocurrido?", se enfoca al igual que el nivel anterior en el pasado a diferencia que esta busca la razón o el motivo de los hechos o patrones que se evidencian en el análisis descriptivo y se requiere aún bastante intervención humana. Nivel Predictivo, busca responder a la pregunta "¿Qué ocurrirá?", se enfoca en el futuro queriendo anticipar los hechos o resultados de forma de tomar acciones de mejora y aun se requiere intervención humana. Nivel Prescriptivo, busca responder a la pregunta "¿Qué debo hacer?", se basa en las predicciones del nivel inferior poder realizar un plan de seguimiento o toma de decisiones de acción.

Teniendo el conocimiento de los diferentes modelos de madurez de los datos tanto en el aspecto de calidad como en el aspecto analítico, en el campo de la salud se han ido desarrollando diferentes modelos para el mejoramiento del manejo de estos volúmenes de datos, tal como se puede apreciar en el trabajo (Carvalho, Rocha, Vasconcelos, & Abreu, 2018) donde se menciona la existencia de modelos de madurez de los datos propuestos para el campo de la salud con la excepción de que aún están en etapa temprana para el desarrollo de un modelo que satisfaga todas las necesidades pertinentes al área de la salud. También se mencionan los diferentes modelos de madurez de los datos que se han presentado en el área de la salud propuestos tales como HIMSS propuestos por los mismos Carvalho, Rocha y Abreu.

Marco teórico

Ciencia de datos

Es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística y la analítica predictiva (López Briega, 2016).

Análisis de datos

El análisis de datos es la ciencia que se encarga de examinar un conjunto de datos con el propósito de sacar conclusiones sobre la información para poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas. Consiste en someter los datos a la realización de operaciones, esto se hace con la finalidad de obtener conclusiones precisas que nos ayudarán a alcanzar objetivos, dichas operaciones no pueden definirse previamente ya que la recolección de datos puede revelar ciertas dificultades (Rouse, 2012).

Limpieza de datos

Es el acto de descubrimiento y corrección o eliminación de registros de datos erróneos de una tabla o base de datos. El proceso de limpieza de datos permite identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc, para luego sustituir, modificar o eliminar estos datos los cuales se conocen como datos sucios (Sánchez Crespo & Villafranca Alberca, 2000).

Messy data

Son datos erróneos o incompletos, especialmente cuando estan en un sistema de ordenadores o en una base de datos. Los datos sucios pueden contener fallos en el deletreo o en los signos de puntuación, datos incorrectos, entre otras cosas (Wickham, 2014).

Visualización de datos

Es la representación fráfica de información y datos. se usan elementos visuales como cuadros, gráficos y mapas, las herramientas de visualizacion de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos (Data Centric, 2017).

Process Enterprise Maturity Model (PEMM)

Este modelo considera dos dimensiones como menciona Mario Saffirio (Saffirio, 2008) los cuales son: Los procesos y la organización.

Procesos: Para los procesos considera como habilitadores de la madurez

- El diseño: propósito, contexto, documentación
- Usuarios: Conocimientos, habilidades, comportamientos frente al cambio
- Dueño: Individualizado, pro-activo, autoridad
- Infraestructura: sistemas de información y recursos humanos

Organizacional: Para la organizacion se considera:

- Liderazgo: alineamiento, comportamiento, estilo
- Cultura: Equipo de trabajo, foco en el cliente, responsabilidad, actitud frente al cambio
- Conocimiento: Personas y metodologías
- Gobernabilidad: Modelos de procesos, contabilidad e integración

Big data

Es un término evolutivo que describe culquier cantidad voluminosa de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraidos para obtener información . Los grandes volúmenes de datos se caracterizan a menudo por 5 Vs: El volumen extremo de datos, la gran Variedad de tipos de datos y la Velocidad a la que se deben procesar los datos (Lara, 2018).

Volumen: El volumen hace referencia al concepto de datos masivos que no se pueden almacenar en un solo ordenador, siendo así necesario la utilización de otras tecnologías para lograrlo.

Variedad: Los datos pueden venir de un archivo, incluyendo datos estructurados, como almacenes de bases de datos SQL. Datos no estructurados, como archivos de documentos o transimisión de datos desde sensores.

Velocidad: Debido a que grandes volúmenes de datos se generan con rapidez, el valor de los mismos se puede perder fácilmente por los nuevos datos entrantes. Por esta razón en un entorno de big data se debe dar una rápida respuesta a esto, teniendo así la necesidad de recopilar, almacenar y procesar estos a gran velocidad.

Veracidad: Se refiere a la calidad de los datos y a que tan eficaces y relacionados entre ellos se encuentran para una toma de decisiones y resultados fiables.

Valor: Este hace referencia a la rentabilidad de los datos, aprovechar los hechos o factores que estos encierran y generar ventajas competitivas. Es el factor mas importante del big Data y es el resultado que agrupa los demás puntos.

Data Mining

Es el proceso de descubrimiento de patrones en grandes conjuntos de datos. Se tiene en cuenta que los conjuntos de datos poseen las características de Big Data. Estos procesos de minería de datos conllevan procesos y análisis exploratorios en grandes volúmenes. El propósito principal es la predicción (predecir con base a datos históricos que se contengan en el conjunto de datos) o como una herramienta de clasificación. La minería de datos (Data Mining) es aplicable en todos los sectores en donde se manejen volúmenes considerables de datos y las ventajas que pueden tener dependiendo de estos (Muñoz, 2017).

Proceso Inmaduro

Es el proceso que tiene total dependencia a los recursos humanos y poca confiabilidad de los resultados debido a la inexactitud de procesos internos, además este no poseer una documentación o registro del funcionamiento, lo que conlleva a que sea difícil anexar a nuevos proyectos (Muñoz, 2017).

Proceso Maduro

Es aquel procedimiento que esta con una descripción clara, detallada, cuenta con sus respectivas entradas, salidas esperadas y establecidas. Está documentado, publicado y es de fácil acceso para reproducir por cualquier trabajador. Además cuenta con una revisión regular donde cada trabajador ha recibido la capacitación pertinente para el uso y práctica de dicho proceso y se mide con base en el rendimiento de éste (Muñoz, 2017).

Metodología

Como fase inicial en el desarrollo del proyecto se dispone a la comprensión del modelo de negocio de la comercialización de productos farmacéuticos y medicinales que realiza la Droguería Puerto Boyacá S.A.S., a su vez con la adquisición y reunión de los conjuntos de datos que manejan en este proceso.

El análisis descriptivo se plantea realizar sobre tres conjuntos de datos; RIPS, medicamentos, y la planilla de entrega de medicamentos. Cada uno de estos conjuntos están asociados o se generan de acuerdo a los contratos establecidos por la droguería con alguna eps o entidad prestadora de servicios de salud.

Conjuntos de datos

RIPS: El Registro Individual de Prestación de Servicios de Salud o RIPS (Ministerio de Salud, 2000), según la resolución 3374 del 2000 es un conjunto de datos mínimos que requiere el Sistema General de Seguridad Social en Salud (SGSSS) para diversos procesos de dirección, regulación y control. En este conjunto refieren datos a la identidad del prestador del servicio, información básica del usuario, datos del servicio de prestación, el motivo que origina la prestación, el diagnóstico y la causa externa.

De forma general este conjunto de datos contiene todos los ciudadanos a atender en un periodo de tiempo (generalmente mensual) con la cantidad máxima del medicamento o insumo hospitalario que se podrá entregar a un ciudadano. Cada una de las entregas realizadas deben ser registradas por la droguería, demostrando así su cumplimiento del contrato.

Inicialmente para dar una utilidad y valor de información a hallar que posteriormente se proporcionará a la Droguería Puerto Boyacá S.A.S. se requiere que los registros de este conjunto específicamente posean datos básicos de los usuarios como identificación, género, edad, lugar de residencia, estrato socioecónomico, medicamento o insumo requerido y su cantidad máxima.

Tabla 1. *Formato esperado para rips*

Id	***Género***	***Edad***	***L.residencia***	***Estrato***	***Medicamento o insumo***	***Presentación***	***Concentración***	***Cantidad***
xxxxxxxx	F	30	xxxxxxxx	2	acetaminofén	tableta o cápsula	500mg	5

Medicamentos: Este conjunto de datos hace referencia a los medicamentos que ha comprado la droguería para satisfacer o cumplir con los contratos y demás ventas particulares. De este se espera que contenga datos relacionados a el nombre del medicamento o insumo, su presentación, concentración, descripción de la compra, nombre del expendio o distribuidor de donde lo compraron, el precio unitario, y por último la cantidad.

Tabla 2. *Formato esperado para medicamentos*

Medicamento o insumo	***Presentación***	***Concentración***	***Descripción***	***Distribuidor***	***Precio***	***Cantidad***
acetaminofén	tableta o cápsula	500mg	caja x 300 tab	distrifarmich	12.000	1

Es pertinente aclarar que las compras de medicamentos o insumos que realiza la Droguería Puerto Boyacá S.A.S. a las distribuidoras de estos son unidades de cajas o paquetes con una determinada cantidad, tal y como se puede ver en la Tabla 2. Para este caso interesa más el análisis descriptivo relacionado al precio por medicamento y distribuidor, verificar la cantidad exacta de los medicamentos actuales en inventario para el cumplimiento de un determinado contrato no es relevante para el propósito del modelo descriptivo ya que los medicamentos que compra la droguería exceden a los necesarios de un contrato en particular, debido a que también se venden medicamentos de forma individual a los ciudadanos, estos osn conocidos como eventos.

Planilla de entrega de medicamentos: La planilla de entrega de medicamentos es el conjunto de datos que se va formando a medida que la droguería realiza una entrega correspondiente a un ciudadano que está en el conjunto de rips. Este registro a pesar de que es la entrega realizada a los ciudadanos generalmente según lo indicado por los responsables del negocio, no hay una correspondencia completa a la cantidad total que se puede realizar.

Los datos presentes aquí corresponden a los mismos en rips con la diferencia anteriormente explicada y sin datos básicos del usuario, además van en conjunto a las fechas de formulación y fecha de entrega, el hospital, y el tipo de médico que formuló.

Tabla 3. *Formato esperado para la planilla de entrega de medicamentos*

Id	***F.formula***	***F.reclamo***	***Hospital***	***Medico***	***Medicamento o insumo***	***Presentación***	***Concentración***	***Cantidad***
xxxxx	dd/mm/aaa	dd/mm/aaa	xxxx	general	acetaminofén	tableta o cápsula	500mg	5

Análisis a realizar

Para llegar al modelo de madurez analítica descriptiva para la organización, se requiere establecer un orden en los conjuntos de datos suministrados por la empresa y determinar el alcance u objetivos del análisis. Para el proceso en cuestión, entonces se pretende la ejecución de tres etapas o fases de análisis; una etapa temprana de limpieza de datos, una intermedia donde se realizará un análisis exploratorio de datos, y finalmente un proceso de llevar los datasets resultantes y pertinentes a una base de datos NoSQL que facilite la consulta y visualización de lo que pasa, es decir, terminar en el estado descriptivo.

1. Etapa de limpieza de datos: En principio el procedimiento que toma lugar es el de limpieza de datos, en el que se tomará cada conjunto para analizar en busca de posibles errores en los datos que impidan o dificulten los análisis posteriores. Primeramente como objetivo se busca normalizar las cadenas de texto dejandolas todas en minúscula y sin tíldes ya que esto hace más díficil los procedimientos siguientes. También se completarán los registros en blanco o nulos por medio de algún método de imputación, además indentificar los datos fuera de lo normal u outliers que sean necesarios eliminar. Finalmente ajustar correctamente los tipos de datos como el categórico y los demás permite ejecutar algoritmos y procesos con mucha más rápidez.

2. Etapa del análisis exploratorio: En este punto se procede a tomar los distintos datasets resultado del anterior para realizar los principales análisis que puedan dar información de valor a la droguería sobre cuál es su situación actual. 1) Principalmente se pretende realizar una caracterización de la población que es atendida mediante diversos criterios como el género, la edad, luegar de residencia, por los medicamentos, y demás que avengan valor e importancia. 2) Se desean hallar además estadísticas descriptivas generales del conjunto de medicamentos que

posee la droguería (este es el conjunto de datos medicamentos). 3) Para terminar, otro análisis al que se plantea llegar es a la comparación de la población máxima que podría atenderse (los ciudadanos y cantidades que se encuentran en rips) con la que realmente está atendiendo (las personas que se encuentran en la planilla de entrega), 4) se podría además conocer el costo total que le está tomando la atención de todos los pacientes que realmente sí reclaman sus medicamentos o insumos.

3. Etapa de migración a base de datos NoSQL: Terminados los análisis de la etapa anterior se planea migrar los conjuntos de datos resultantes a MongoDB (MongDB, s.f.) una base de datos NoSQL distribuida y basada en documentos diseñada especialmente para la era de la nube, la cual posee ventajas como (GAD, s.f.): mejores resultados al momento de manejar grandes cantidades de documentos, costo bajo, puede realizar muchas operaciones por segundo, y es de fácil escalabilidad. Contrastando esto a la situación de una mipyme como la Droguería Puerto Boyacá S.A.S. se ajusta bastante bien debido a tres situaciones que se presenta; 1) la mala administración de datos (o ausencia de esta) que se suele presentar en este tipo de empresas genera documentos con registros incompletos, erroneos o de distintos formatos, lo que en mongodb se facilita y ayuda a diversos análisis; 2) las pymes o mipymes no cuentan con mucha financiación o presupuesto; y finalmente 3) gracias a la fácil escalabilidad de mongodb por su distribución en clúster facilitará a futuro la expansión del negocio.

Para este propósito se hará uso de MongoDB Atlas (MongoDB, s.f.) un servicio de base de datos en la nube que ofrece varias ventajas en su fácil administración, escalabilidad, seguridad y privacidad. Además en está se ofrece una versión gratuita con capacidad de hasta 500 MB, cantidad de datos que para los inicios de una mipyme y para los propósitos de este proyecto es suficiente.

Una vez migrados los documentos respectivos se procederá a la construcción de algunas consultas básicas que se asemejen a los análisis realizados, dejando así implementado un modelo que le posibilite a la droguería la visualización de su estado actual. Para esto se dispondrá de MongoDB Compass (MongoDB, s.f.) un software libre que facilita la exploración, visualización e interacción con los datos, así como la facilidad de ejcutar queries sencillamente y en pocos segundos.

Proceso de limpieza de datos

La primera anotación a realizar es que se manejan distintos formatos de archivos en los que actualmente son almacenados los datos, siendo archivos de excel en su mayoría, pdfs y algunos otros datos se encuentran en imágenes (jpeg). Debido a este problema no se pudieron tomar todos los datos necesarios para un análisis completo, si bien los archivos de excel no son los más indicados estos es posible tomarlos y procesarlos, sin embargo, para los archivos pdf e imágenes se debe transferir está información para lo cual es necesario por parte de la empresa realizar una digitación de la misma.

Es comunmente dicho que el 80% del análisis de datos se encuentra en el proceso de limpieza y preparación de los datos (Tanraparni & Theodore, 2003) y en este caso dicho trabajo tuvo lugar con el procesamiento de los archivos de excel mencionado anteriormente. Estos represtan los tres conjuntos de datos; rips, planilla de entrega de medicamentos, y medicamentos, donde cada uno fue procesado para el posterior análisis.

El primer procedimiento realizado fue la revisión del cumplimiento del concepto de tidy data (Wickham, 2014) sobre los distintos archivos suministrados, específicamente sobre los primeros dos criterios del tidy data:

- Cada columna debe representar una variable.
- Cada fila debe representar una observación.

Con el cumplimiento de estos dos criterios el análisis es más eficiente, además el proceso de limpieza siguiente se facilita. Lo hallado fue que cada archivo en los formatos de tabla que trabaja la Droguería Puerto Boyacá S.A.S. cumple perfectamente con los criterios.

Durante esta revisión sobre los conjuntos de datos se comprobó una inconsistencia entre la cantidad de tablas o meses del archivo rips y el archivo de planilla de entrega de medicamentos, en donde el primero solo contaba con datos de dos meses (que fueron asumidos como enero y febrero) y el segundo con datos de 7 meses (de enero a julio).

Para el análisis deseado además del preprocesamiento y limpieza de datos era necesario una correspondencia por meses en los archivos de rips y planilla de entregra de medicamentos, ya que este primero contiene el limite de posibles entregas que deberían como máximo estar registradas en el segundo. Con esto se deseaba comprobar el hecho de ganancia o pérdida de la ejecución del contrato en cuestión.

Debido al problema anterior mencionado del cual se desconoce la causa de la falta de la información necesaria, se tomaron solo los primeros dos meses para la realización del análisis. Se ha de destacar que este hecho no afecta al archivo o conjunto de datos de medicamentos el cual es independiente de los otros dos, logrando así realizar una caracterización y análisis descriptivo de forma general.

Se procede ahora a mostrar ahora el proceso de limpieza de datos realizado:

Limpieza de datos en la planilla de entrega de medicamentos

Este archivo fue el que más problemas tuvo, algunos de estos impidiendo la posibilidad de llegar a algunos análisis planteados anteriormente, y otros son problemas comunes de la limpieza de datos como: revisión de outliers o valores atípicos, parseo de fechas, e imputación de valores nulos o faltantes (Wickham, 2014). Se ha de mencionar que en el mes de Febrero no se hallaron tantos problema como en la tabla correspondiente a Enero, por lo que sólo se mostraran los hallazgos de este último.

Estos problemas se listan a continuación:

Información básica de las personas

La información de los usuarios se encontró en este archivo en lugar del de rips. Esto a pesar de no ser un problema de la limpieza de datos fue un hecho que impidió más adelante hacer una comparación entre la caracterización de la población objetivo, es decir, aquella que se encuentra en rips, y la caracterización de la población realmente atendida, la que se encuentra en este archivo. Esto ocurre puesto que no todas las personas o números de identificación personal que se encuentran en la planilla estan en el archivo de rips.

Formatos de tablas diferentes

Los primeros tres meses; enero, febrero y marzo, contienen columnas de datos para toda la información básica de las personas, sin embargo, los meses restantes no contienen dichas columnas lo que impide como en el problema anterior poder realizar una caracterización de toda la población que atiende, debido a que no son iguales todos los usuarios atendidos entre mes y mes faltando así información de algunas personas.

Inconsistencia en nombres de columnas

Este problema a pesar de no impedir la ejecución de los análisis sí dificulta el proceso, por tal rezón es recomendable normalizar estos nombres.

Tabla 4. *Formato planilla de entrega de medicamentos recibido*

HOSPITAL	***MEDICO***	***FECHA***	***TiIPO***	***IDENTIFICACION***	***SEXO***	***EDAD***	***PRICIPIO ACTIVO***	***CONCENTRACIÓN***	***FORMA FARMACÉUTICA***	***CANTIDAD***
XXXXX	XXXX	x/x/x	CC	xxxxxxx	M	40	TIAMINA	300mg	TABLETA	60

Se omiten columnas de información personal como nombres y teléfono, además de algunas con códigos de manejo interno para los medicamentos.

Como se puede ver en el formato de la planilla de entrega de medicamentos de la tabla 4 los nombres de columnas se encuentran en mayúscula, alguno con tíldes y espacios, además la columna fecha en el archivo original de excel se encuentra agrupada en dos; formula y reclamo, lo cual al leer el archivo genera problemas.

Estos problemas fueron solucionados normalizando todos los nombres a minúscula, sin tíldes, cambiando espacios por guión bajo y quitando las columnas que no fuesen necesarias (de información personal y manejo interno).

Valores nulos o faltantes

La tabla correspondiente al mes de Enero presenta un total de 1666 registros o filas, a continuación se procede a mostrar las cantidad de valores presentes en cada columna junto al tipo de dato de la misma.

Tabla 5. *Cantidad de valores presentes en la planilla de entrega para el mes de Enero*

Columna	***Valores presentes***	***Tipo de dato***
hospital	515	object
medico	515	object
fecha_formula	1663	object
fecha_reclamo	1663	object
tipo	1663	object
identificacion	1663	float64
sexo	1615	object
edad	516	float64
principio_activo	1661	object
concentracion	1658	object
presentacion	1660	object
cantidad	1652	object

Los tipos de datos mostrados son los manejados por la librería pandas.

En la tabla 5 se evidencia la bastante cantidad de valores faltantes para las columnas hospital, medico y edad, mientras el resto de columnas realmente estan muy completas. Para este problema se procedió a realizar un revisión de los valores presentes en cada columna para decidir así un buen mecanismo de imputación.

Para las columnas de hospital y medico se encontro un valor único en cada columna; H. VASQUEZ y GENERAL respectivamente. Al ser un valor único este no afecta el análisis y por tal razón se decidió quitar estas dos columnas.

En la revisión del resto de columnas se pudo evidenciar una tendencia por omitir la redigitación de la misma información referente a una persona, es decir, si en una fila se encuentra toda la información de esta persona en la que sigue solo colocaron los campos que cambiaban y dejando siempre el número de identificación. Por esto para el mecanismo de imputación en el resto del dataset se hizó uso del metódo del dataframe en pandas (pandas, 2011) *fillna* con su argumento *method* en *ffill*, el cual establece que se imputará de forma ascente con el registro inmediatamente anterior al de la posición actual.

Tipos de datos incorrectos

En la tabla 5 se encuentran las columnas con el tipo de dato obtenido al leer el archivo, cada una de las columnas tuvo un respectivo casteo para mejorar la eficiencia y poder realizar un correcto análisis.

Para las columnas de fecha_formula y fecha_reclamo se realizó un casteo a el tipo de dato datetime64 manejado por pandas, el cual permite la ejecución de análisis con series de tiempo y otras operaciones con los datos tipo fecha. En este proceso se hallaron problemas en algunas fechas digitadas erroneamente; algunas con doble slash (//), otras con guiones en lugar de slashes, algunas otras contenian solo los dos últimos digitos del año, y demás, al realizar el casteo todas estas fechas se convirtieron en valores nulos que tuvieron que ser imputados como en el paso anterior.

Para la columna identificación se realizó un casteo al tipo de dato object-

La columna cantidad fue casteada al tipo de dato int64, el cual es el indicado para poder luego realizar una revisión de outliers y otros análisis. Los valores nulos resultantes de la imputación fueron procesados al igual que con las fechas.

El restante de columnas fueron casteadas a un tipo de dato categórico, el cual permite la ejecución de análisis más rápido y eficiente, disminuyendo el consumo de memoria y espacio.

Revisión de outliers

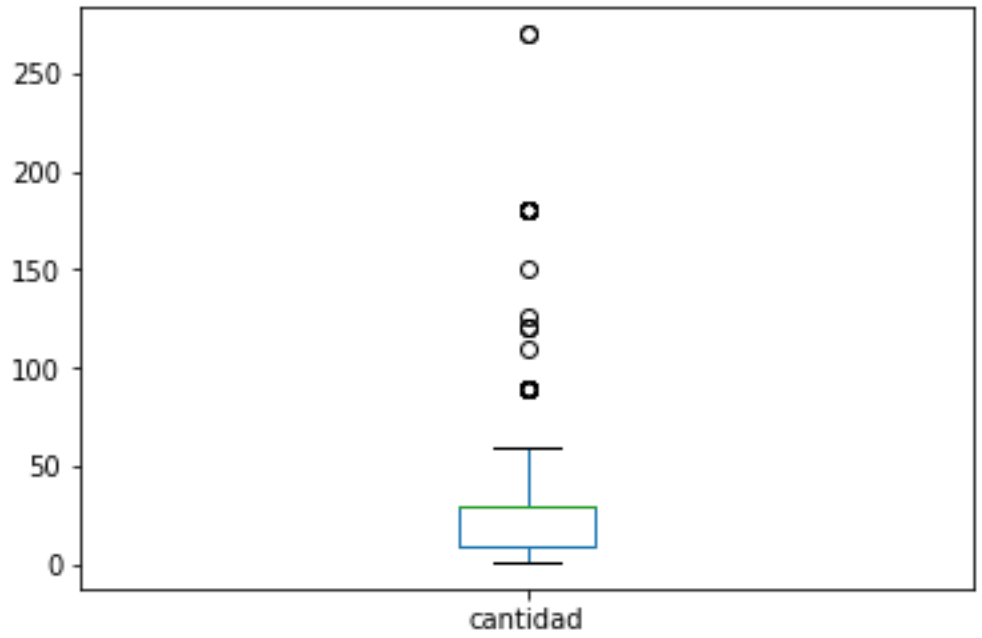

Ilustración 2. Boxplot para la columna cantidad del mes de Enero

Fuente: Elaboración propia

En la ilustración 2 se puede ver el boxplot de la columna cantidad, el cual muestra además del promedio y los limites según a los valores más comunes, los valores considerados como outliers o atípicos. En esta se nota que la mayor cantidad de entrega de medicamentos a los usuarios se encuentra entre 1 y 70 aproximadamente, y otros valores superiores a 100 y 200 que son atípicos. Para estos valores se procedió a realizar una comparación con la cantidad de entregas máximas en rips mes 1 para verificar su relación.

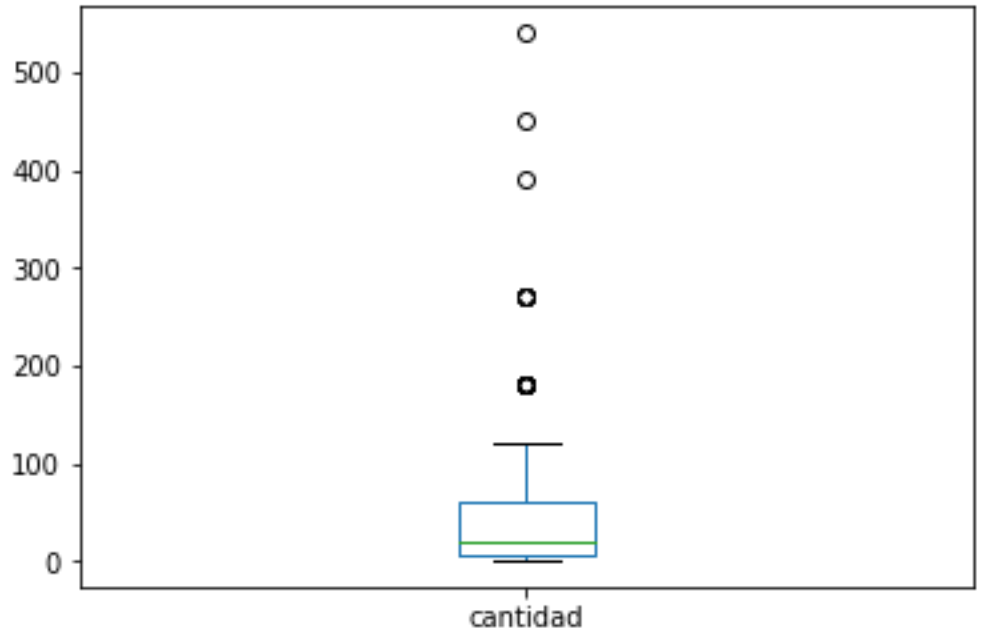

Ilustración 3. Boxplot columna cantidad mes 1 de rips

Fuente: Elaboración propia

Realizando la comparación entre el boxplot de enero y el boxplot del mes 1 de rips se evidencia una similitud en su promedio y limites, asimismo los outliers del mes de enero que superan la cantidad de 100 y 200 se encuentran también en rips, por lo que estos no son valores erróneos y se dejan en el conjunto de datos.

Otra anotación a realizar es que en el mes 1 de rips se encuentran otros outliers mucho más altos estando por encima de 300, 400 y 500, de los cuales por parte de la información suministrada por la Droguería Puerto Boyacá S.A.S. se encuentran bien, por lo que también se dejan en el dataset.

Limpieza de datos en el archivo de rips

El archivo de excel proporcionado por la droguería se encontraba dividido en dos hojas llamadas como mes 1 y mes 2. Estas al comparar sus registros con los meses de Enero y Febrero en la planilla de entrega de medicamentos se pudó comprobar bastante similitud en los números

de identificación, por lo que se asumió que eran las equivalentes para los rips, siendo mes 1 los rips para Enero y mes 2 los rips para Febrero. Ambos conjuntos de datos se encontraban bastante bien, llegando así a no tener que hacer casi nada como preprocemamiento o limpieza de datos.

Como fue mencionado en el punto anterior de la planilla de entrega de medicamentos estas hojas no contenían la información básica de los ciudadanos. El formato recibido fue:

Tabla 6. *Formato recibido para las tablas de rips*

EPS	***DOC***	***CC***	***CODIGO***	***MEDICAMENTO***	***PRESENTACION***			***CANTIDAD***
xxxx	CC	xxx	xxxxx	LOSARTÁN	TABLETA O CÁPSULA	50mg	MILIGRAMOS	120

Las dos columnas sin nombre corresponden a la concentración y unidad de medida, la columna eps corresponde a un código identificador de la eps del contrato en cuestión, y por útimo la columna codigo es de manejo interno para específicar determinado medicamento.

Nombres de columnas

El principal proceso de limpieza de datos llevado acabo sobre estos conjuntos fue ajustar los nombres de columnas, en donde se normalizaron como en las tablas respectivas de la planilla de entrega de medicamentos. Además de este dataset se eliminaron las columnas de EPS; la cual es siempre el mismo valor, la columna de CODIGO; esta contiene códigos dados a los medicamentos, a pesar de que podría ser más eficaz trabajar con estos, no fue posible ya que no se encontraban en todas las demás tablas de la planilla de entrega ni en el dataset general de medicamentos, y por último la columna sin nombre correspondiente a unidad de medida fue eliminada porque en la propia columna de concentración se encuentra esta unidad.

Revisión de outliers

En las hojas de rips no se encontraron valores faltantes por lo que no se requirio ningún proceso de imputación. Sin embargo, sí se realizó una revisión de outliers en ambos conjuntos. Los outliers de la hoja mes 1 se evidencian en la ilustración 3, así que se procede a mostrar los mismos para el mes 2.

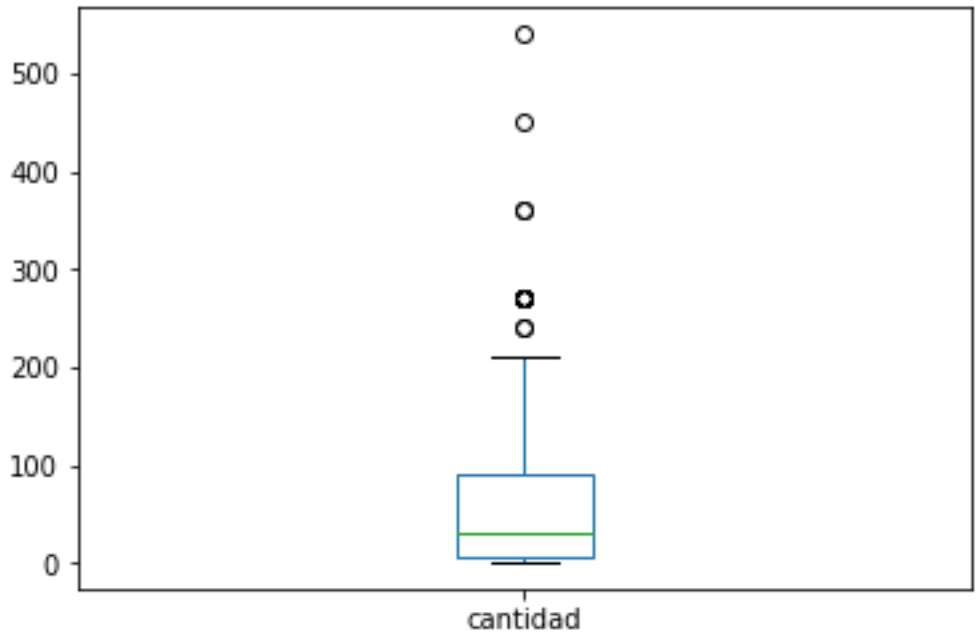

Ilustración 4. Boxplot columna cantidad mes 2 de rips

Fuente: Elaboración propia

Como se puede observar en esta ilustración, es similar al boxplot del mes 1, donde se encuentran limites muy parecidos, y se tienen además casi que los mismos outliers que superan una cantidad de 300, 400 y 500.

Con la revisión de estos valores atipícos se toma la decisión de dejarlos así debido a que según lo dictado por la droguería; no están mal y pertenecen a lo establecido de posibles entregas a realizar.

Limpieza de datos para el archivo de medicamentos

Sobre este conjunto no se realizaron muchas actividades de limpieza de datos, principalmente el trabajo hecho fue ajustar el dataset ya que en las hojas de excel usadas por la Droguería Puerto Boyacá S.A.S. se encuentran varias tablas en una misma a pesar de que estas tengan el mismo formato o columnas, esto es desorganización en la estructura manejada, por lo que fue necesario ajustar la estructura y dejar todos los datos en una sola tabla.

Los nombres de columnas utilizados no fueron modificados ya que aunque estos no estuviesen en minúscula como el estándar dejado para los datasets restantes, estos sen econtraban todos en mayúscula (sin mezclar minúsculas como en los casos anteriores), sin tíldes y sin espacios, lo que posibilita trabajar sin dificultades extra.

Y por último, para las demás actividades de limpieza de datos como la imputación y revisión de outliers no fueron necesarias, gracias a que inicialmente no se hayaron valores nulos o faltantes y por otra parte no destacó ningún valor atípico, todas las cantidades y precios se encontraban bien. El único problema hallada fue el formato de tabla recibido.

Distinto formato de tabla

El formato de tabla que usa la droguería para la específicación de los medicamentos es distinto al que se ha establecido en los demás conjuntos, dificultando así realizar algunos análisis.

Tabla 7. *Formato recibido para el conjuto de datos de medicamentos*

COD	***DESCRIPCION***	***PRESE***	***LABORATORIO***	***IVA***	***CANT***	***PRECIO***	***TOTAL***	***DESC***	***LOTE***	***F.VENCIM***
XXX	ACETAMINOFEN 500MG X300 TAB	CAJA X 300 TABLETAS	XXXX		1	10.000	10.000			

Como se puede observar y en comparación al formato esperado, este aunque contiene los datos necesarios es algo diferente. La columna COD es un código (este es distinto al usado en las otras tablas), y las columnas DESC, LOTE y F.VENCIM se encontraban todas completamente vacías. Para el caso de la columna LABORATORIO esta contiene el fabricante del medicamento mas no es distribuidor al que la droguería realizó la compra, cada distribuidor al cual son comprados los medicamentos se encuentran en hojas diferentes en el archivo de excel por lo que fue necesario juntar estas dos hojas en una sola tabla.

En las otras columnas donde se encuentra la información referida al medicamento es donde realmente estuvo el principal problema, debido a que se esperaba obtener el medicamento separado de su presentación y concentración, sin embargo, en la columna DESCRIPCION se encuentran estos datos juntos, para este caso no decidió separar la información en distintas columnas como en el formato esperado, sino al momento de realizar el análisis aplicar distintos filtros y expresiones regulares para obtener los datos deseados.

Análisis exploratorio de datos

Caracterización por género en la planilla de entrega

Tabla 8. *Cantidad de hombres y mujeres atendidos en el mes de Febrero*

Sexo	***Atendidos***
Mujeres	890
Hombres	458

Para el análisis exploratorio del mes de febrero se cuentan con un total de 1364 registros, es decir 1364 entregas realizadas a los pacientes que se pueden evidenciar en la ilustracion 5 de los cuales 890 son mujeres y 458 son hombres y 16 registros de entregas que no presentan el sexo de la persona que hizo recibimiento de dichos medicamentos, por lo tanto estos no se tuvieron en cuenta para el análisis exploratorio y caracterización de la población. Este se realizo con el fin de visualizar el flujo de entrega de medicamentos.

También se realiza una caracterización con base en el tipo de documento en el cual se aprecia la edad promedio de los pacientes que van a reclamar sus medicamentos. Estos datos se pueden apreciar en la siguiente tabla.

Tabla 9. *Top 10 medicamentos entregados a hombres en el mes de Febrero*

Principio activo y concentración	***Entregados***

Losartán Potásico 50 mg	1170
Acetaminofén 500 mg	858
Acetíl Salicílico Ácido 100 mg	690
Amlodipino 5mg	630
Omeprazol 20 mg	570
Valproico Ácido 250 mg	420
Atorvastatina 40 mg	390
Hidroclorotiazida 25 mg	390
Naproxeno 250 mg	382
Nifedipina 30 mg	330

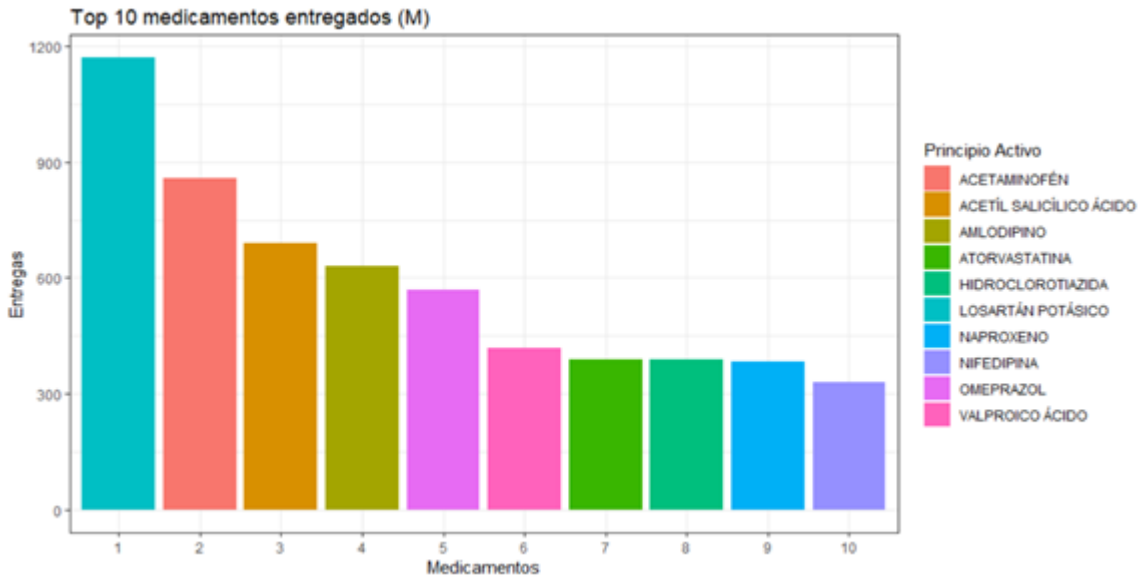

Ilustración 5. Top 10 medicamentos entregados a hombres en el mes de Febrero

Fuente: Elaboración propia

Según los datos proporcionados en "planilla entrega de medicamentos" para el mes de febrero, de las 458 registros de entregas que se realizaron a pacientes de sexo masculino, se pueden observar el top 10 de los medicamentos que más reclamaron y se entregaron por parte de la Drogueria Puerto Boyaca S.A.S. en la tabla 9 con su respectiva gráfica en la ilustración 5.

Tabla 10. *Top 10 medicamentos entregados a mujeres en el mes de Febrero*

Principio activo y concentración	***Entregados***
Losartán Potásico 50 mg	2400
Acetaminofén 500 mg	2051
Acetíl Salicílico Ácido 100 mg	1830
Amlodipino 5mg	1110

Enalapril Maleato 20 mg	870
Omeprazol 20 mg	830
Hidroclorotiazida 25 mg	780
Atorvastatina 20 mg	690
Esomeprazol 20 mg	630
Naproxeno 250 mg	607

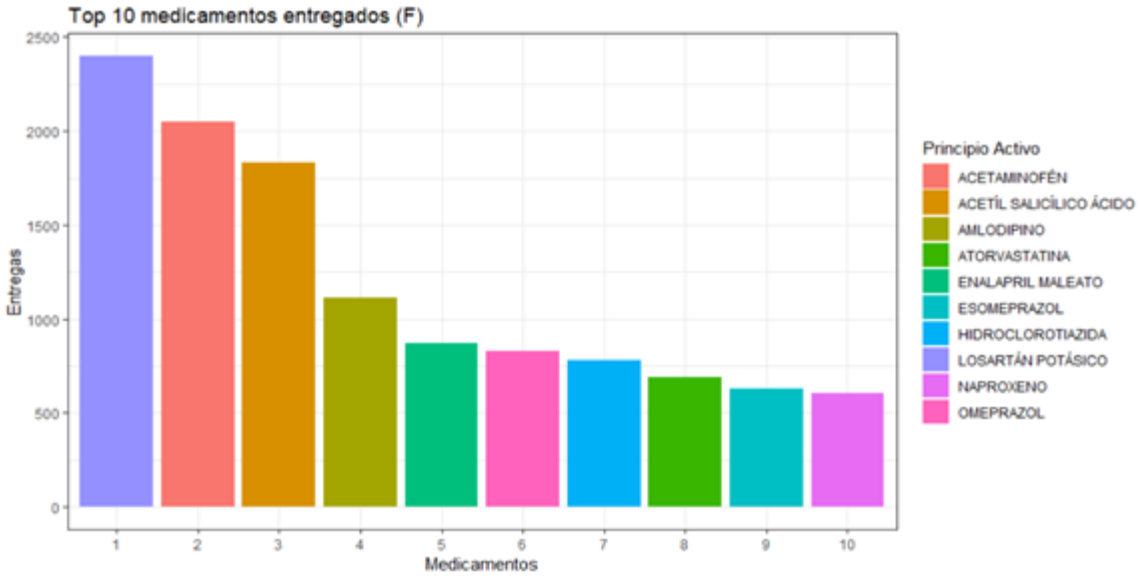

Ilustración 6. Top 10 entrega de medicamentos a mujeres en el mes de Febrero

Fuente: Elaboración propia

Caracterización por tipo de documento en la planilla de entrega

Tabla 11. *Caracterización por tipo de documento*

Tipo de documento	***Edad promedio***
CC	61
TI	14
RC	3

Teniendo en cuenta el número de entregas de medicamentos y la edad promedio, se procede a analizar el top 10 de los medicamentos que mas reclamaron tanto los hombres como mujeres en el mes de febrero.

Caracterización por medicamento

Tabla 12. *Top 10 medicamentos que más se entregaron en el mes de Febrero*

Principio activo y concentración	***Entregados***
Losartán Potásico 50 mg	3570
Acetaminofén 500 mg	2909
Acetíl Salicílico Ácido 100 mg	2520

Amlodipino 5mg	1740
Omeprazol 20 mg	1400
Enalapril Maleato	1170
Hidroclorotiazida 25 mg	1170
Naproxeno 250 mg	989
Atorvastatina 20 mg	810
Metformina 850 mg	810

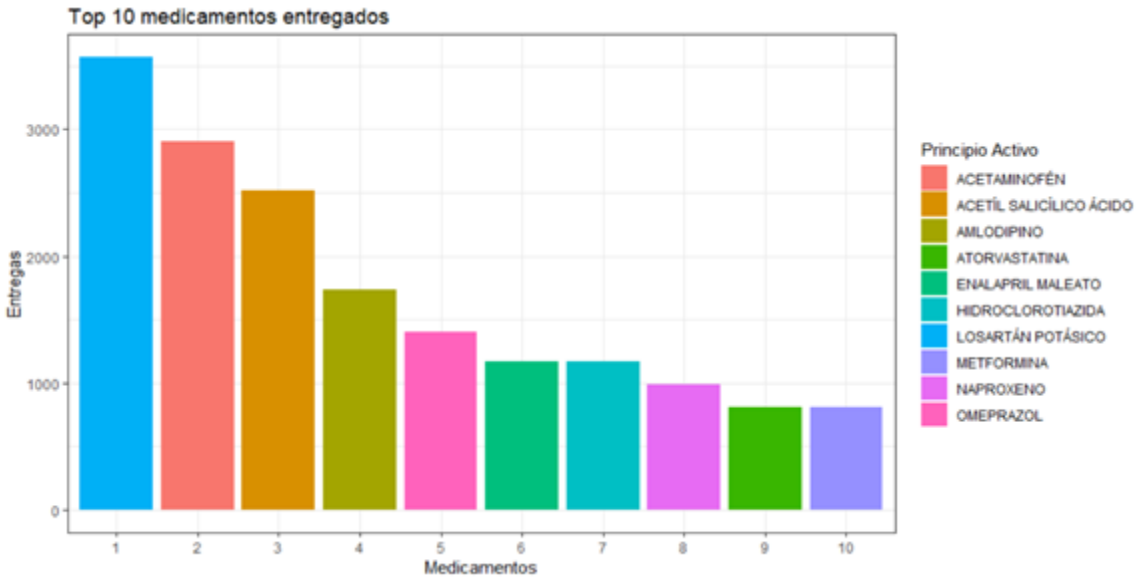

Ilustración 7. Top 10 medicamentos más entregados en Febrero

Fuente: Elaboración propia

Como se observa en ambas tablas e ilustraciones, hubo mayor reclamo de medicamentos por parte de mujeres que de hombres, además se puede apreciar que ambos sexos reclamaron los mismos 4 medicamentos del top en el mes de febrero. Con base en esto se procede a hacer un análisis comparativo con los datos de rips.

Análisis comparativo entre la planilla de entrega de medicamentos y rips

Tabla 13. *Tabla informativa de los top 10 de medicamentos y los rips para mes de Febrero*

Principio activo y concentración	***Entregados***	***Totalrips***	***Porcentaje***	***Precio Neto***	***Total***
Losartán Potásico 50 mg	3570	9150	39.02%	27.00	96390
Acetaminofén 500 mg	2909	5014	59.21%	212.52	618220.7
Acetíl Salicílico Ácido 100 mg	2520	5730	43.97%	16.00	40320
Amlodipino 5mg	1740	2670	65.16%	15.00	26100
Omeprazol 20 mg	1400	3770	37.13%	55.00	77000
Enalapril Maleato	1170	2850	41.05%	- -	- -
Hidroclorotiazida 25 mg	1170	2550	45.88%	- -	- -

Naproxeno 250 mg	989	1028	96.20%	- -	- -
Atorvastatina 20 mg	810	1890	42.85%	- -	- -
Metformina 850 mg	810	2880	28.12%	- -	- -

En esta tabla comparativa se hace un contraste entre el top10 de medicamentos entregados en el mes de Febrero junto con los medicamentos que se establecen en los rips para este mes (basándonos en la hipótesis que mes 2 en el documento de rips corresponde al mes de Febrero debido a los datos relacionados). Es necesario recalcar que para este análisis no se tomó en cuenta las 16 entregas que no poseen el valor sexo con el cual se ha establecido la mayor parte de este análisis exploratorio debido al desconocimiento de estos y la poca información que se poseen de estas entregas. Por lo tanto, esta comparación se hace con la información establecida de los tops de medicamentos entregados a hombres y a mujeres que salieron del documento de plantilla de entrega de medicamentos y de análisis previos.

En el análisis entre los tops de medicamentos entregados y la información suministrada por los rips se ha establecido una comparación del número de medicamentos a entregar según el documento de los rips con el número de medicamentos entregados en todo el mes de Febrero, esto se puede apreciar en la columna porcentaje la cual muestra el comportamiento de los datos con base en lo que se entregó con lo destinado a entregar evidenciando de que, por lo menos en este mes, no hubo una demanda mayor a lo establecido en el documento de los rips y se puede ver que el medicamento que estuvo cerca a llegar a su límite de entrega fue naproxeno 250 mg con un porcentaje de 96.20% .

Otro documento que se tuvo en cuenta para este análisis comparativo fué el de medicamentos donde se posee la información de cada uno de estos que fueron suministrados a la Drogueria Puerto Boyacá S.A.S así como se puede apreciar en la tabla 3. Este se usó con el fin de sacar el precio neto de cada uno de los medicamentos que forman parte del top 10 de entregados pero sólo se contó con la información de los 5 medicamentos que encabezan el top, de los cuales los 5 restantes no presentan información alguna en el documento de medicamentos suministrado lo que imposibilita el análisis de valor total invertido por la droguería.

Migración a base de datos NoSQL

Uno de los principales problemas del porque la mayoría de pymes y mipymes se quiebran en el primer año es por la falta de financiación y una mal toma de decisiones (COLOMBIA FINTECH, 2019) hecha sin buenas bases o fuentes de información clara. Esta toma de decisiones críticas puede mejorar si la organización avanza en madurez analítica de la información.

Se decidió por migrar los datos MongoDB ya que este motor de base de datos open source que por sus características, herramientas y usos da solución a los problemas mencionados. En su versión MongoDB Atlas, un servicio de base de datos en la nube, se ofrece una versión gratuita que a pesar de tener solo 500 MB de almacenamiento es una cantidad que soporta bastante información que generan las mipymes como la Droguería Puerto Boyacá S.A.S. permitiendo así ahorrar costos de acomodación de espacio y manutención de un servidor físico. Ahora bien, con la herramienta MongoDB Compass, también open source, la cual ofrece diversos análisis descriptivos y visualizaciones sin mayor esfuerzo, permitiendo además llegar a otros análisis más complejos facilmente.

Gracias a estas herramientas solo es necesario por parte de la empresa definir dos elementos más que le permitan llegar a los análisis deseados en un nivel de madurez analítico descriptivo, estos son; ajustar o definir las estructuras de documentos con los datos mínimos a ser procesados en conjunto a los análisis o información que se quiere obtener, y por otra parte tener personal especializado para estas labores.

Resultados obtenidos en MongoDB Compass

El primer beneficio notorio al usar MongoDB Compass es el schema, una opción que ofrece estadísticas y gráficas básicas que nos permiten identificar la estructura general de un determinado conjunto de datos. Permitiendo observar los tipos de datos presentes en cada columna y asímismo los valores que integran dicha columna. Se ha de aclarar que esta información dada en el schema se obtiene al analizar una muestra de 1.000 registros o documentos (documento es el término usado en mongo) de la colección en cuestión.

A continuación se muestran algunas capturas de pantalla tomadas sobre el conjunto de datos o colección (colección es el término usado para indicar un conjunto de datos) de entregas del mes de Febrero.

Cantidad de entregas

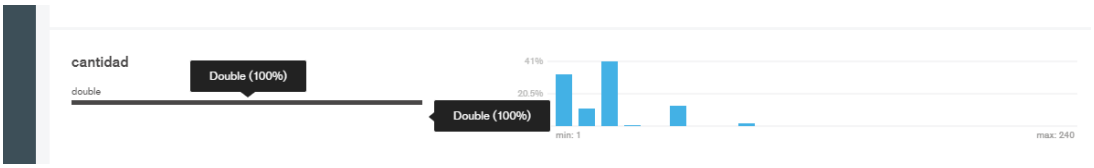

Ilustración 8. Tipos de datos para la columna cantidad dados por MongoDB Compass

Fuente: Elaboración propia

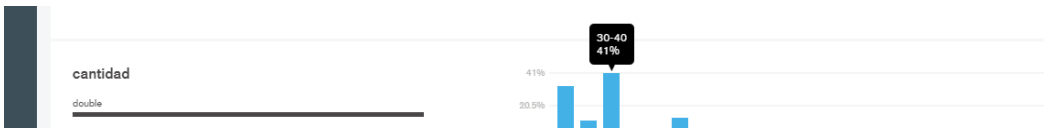

Ilustración 9. Cantidad de entregas en el schema de MongoDB Compass

Fuente: Elaboración propia

En la ilustración 8 y 9 se pueden observar los intervalos dados por Compass, en estas se observa primeramente que el total de los datos estan identificados como double (el 100%) y en los intervalos específicamente sobre el que se encuentra el mouse el cual es de 30 a 40 tiene un total del 41% de los datos.

Edad

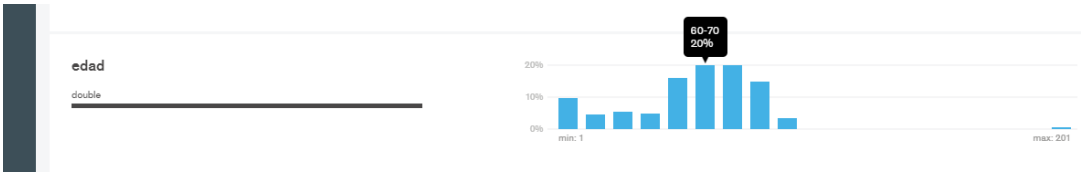

Ilustración 10. Intervalos de edad dados por Mongo DB Compass

Fuente: Elaboración propia

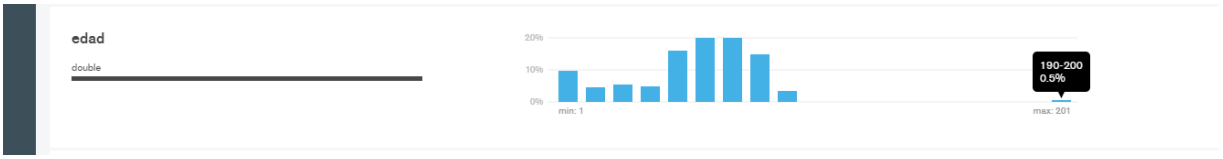

Ilustración 11. Outliers en edad identificados en Mongo DB Compass

Fuente: Elaboración propia

En la ilustración 10 se observan los intervalos de edad dados en el schema, en esta se observa que el intervalo entre 60 y 70 años junto al que le sigue (este es de 70 a 80 años) son los más representativos teniendo un 20% del total de los datos cada uno.

En la ilustración 11 se puede observar un outlier en edad, el cual es un intervalo de entre 190 a 200 años. Se puede apreciar entonces como Mongo DB Compass también nos puede ayudar a identificar facilmente problemas comunes para la limpieza de datos; como la revisión de outliers y los tipos de dato.

Sexo y tipo de documento

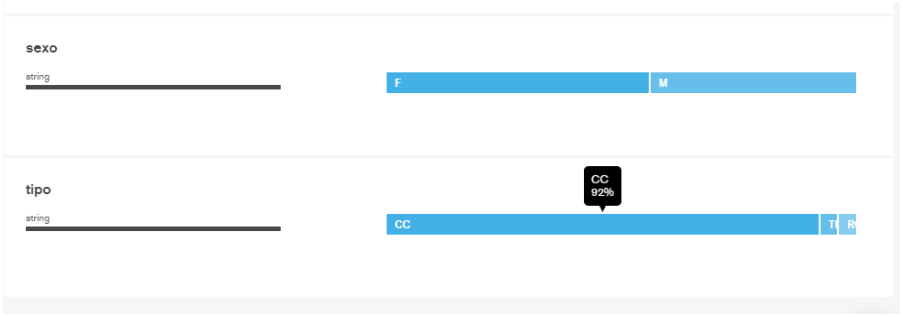

Ilustración 12. Información del sexo y tipo de documento en MongoDB Compass

Fuente: Elaboración propia

En la ilustración 9 se evidencia el tipo de dato para las variables “sexo” y “tipo”, además de poder ver de forma rápida los datos que conforman o hay presentes en estas variables tal como es el caso de la variable tipo donde presenta 3 tipos de datos que son CC, TI y RC donde la cédula de ciudadania representa un 92% de los datos que MongoDB Compass usó de muestra para el schema.

Agregaciones

Las operaciónes de agregación procesan registros de datos y devuelven resultados calculados. Las operaciones de agregación agrupan valores de varios documentos y pueden realizar una variedad de operaciones en los datos agrupados para obtener un único resultado (MondoDB inc, 2008). Se realizó una agregación para poder sacar los tops que se sacaron previamente y poder visualizarse en MongoDB Compass.

Match

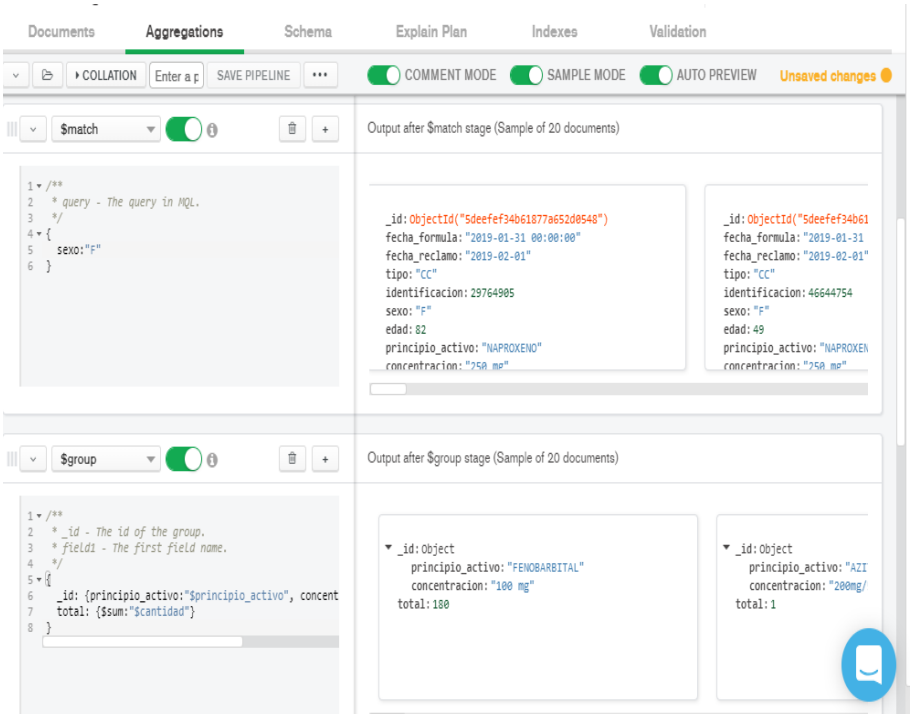

Ilustración 13. Agregación match

Fuente: Elaboración propia

En esta primera agregación en MongoDB Compass lo que se hace es filtrar los datos que cumplan con la condición de que el sexo sea igual a “F” para tener los registros de las entregas a pacientes de sexo femenino.

Group

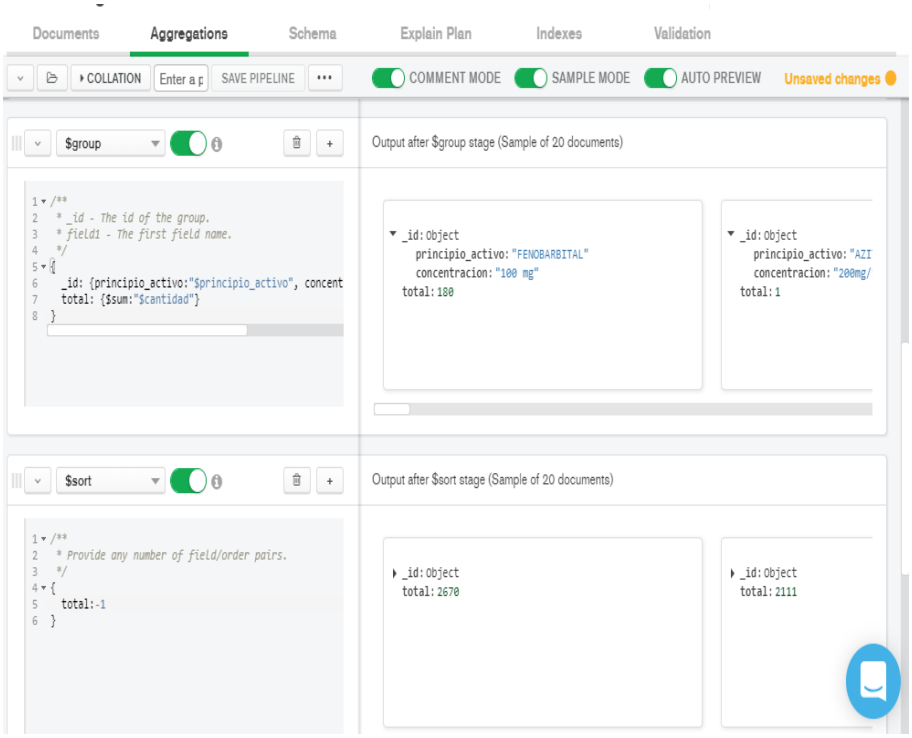

Ilustración 14. Agregación group

Fuente: Elaboración propia

En esta segunda agregación lo que se quiere hacer es agrupar los datos con base en el principio activo y la concentración con el fin de determinar la cantidad total de dichos medicamentos que fueron entregados en el mes de febrero.

Sort

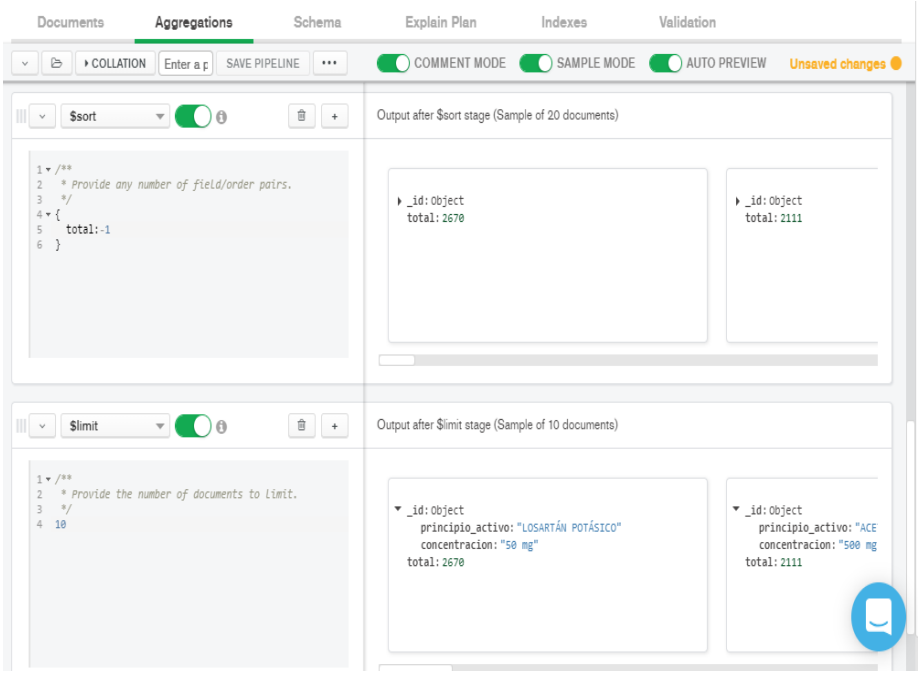

Ilustración 15. Agregación sort

Fuente: Elaboración propia

En esta tercera agregación sort lo que se quiere hacer es que se reordene la consulta de forma descendente, es decir de mayor a menor con base al total que dio como resultado la agregación anterior group.

Limit

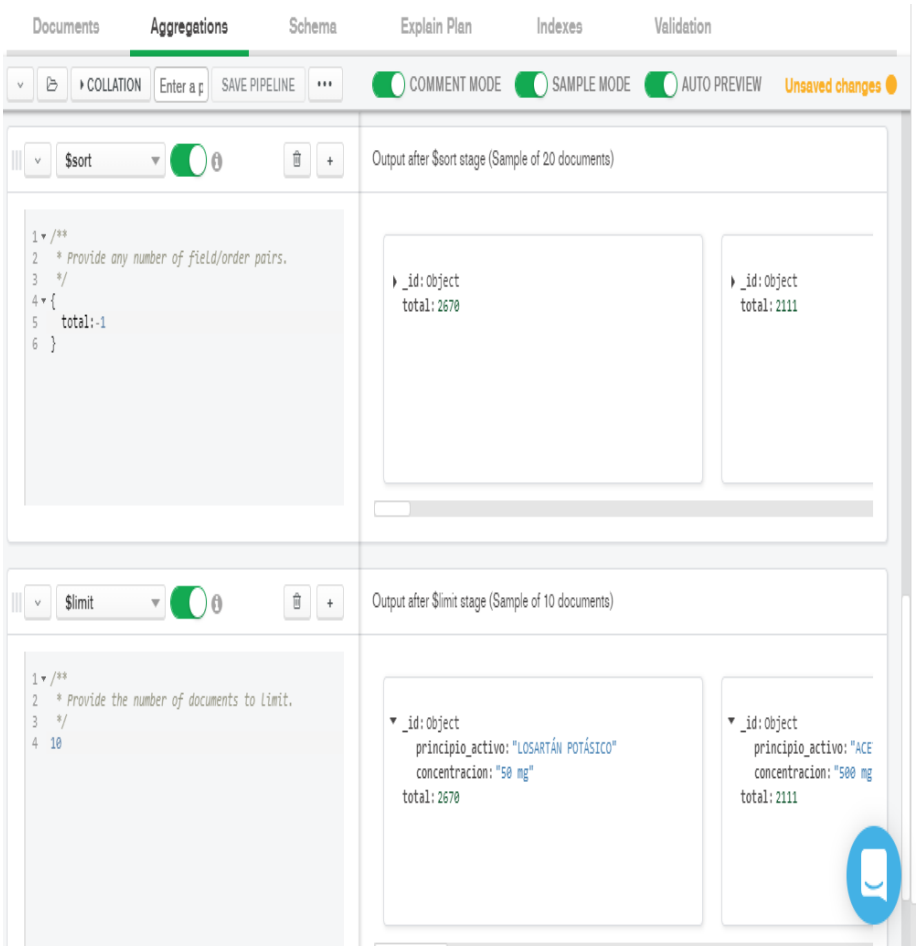

Ilustración 16. Agregación limit

Fuente: Elaboración propia

En esta última agregación limit lo que se quiere con esta es que solo retorne un número maximo de resultados que uno designe tal como fue para este caso 10.

Consulta desde línea de comando

```
MongoDB Enterprise drogueria-shard-0:PRIMARY> db.entregas.aggregate([{$match:{sexo:"F"}}, {$group: {_id:{principio_activ
o:"$principio_activo", concentracion:"$concentracion"}, total:{$sum:"$cantidad"}}}, {$sort:{total:-1}}, {$limit:10}])
{ "_id" : { "principio_activo" : "LOSARTÁN POTÁSICO", "concentracion" : "50 mg" }, "total" : 2670 }
{ "_id" : { "principio_activo" : "ACETAMINOFÉN", "concentracion" : "500 mg" }, "total" : 2111 }
{ "_id" : { "principio_activo" : "ACETÍL SALICÍLICO ÁCIDO", "concentracion" : "100 mg" }, "total" : 1830 }
{ "_id" : { "principio_activo" : "AMLODIPINO", "concentracion" : "5 mg" }, "total" : 1110 }
{ "_id" : { "principio_activo" : "ENALAPRIL MALEATO", "concentracion" : "20 mg" }, "total" : 990 }
{ "_id" : { "principio_activo" : "OMEPRAZOL", "concentracion" : "20mg" }, "total" : 860 }
{ "_id" : { "principio_activo" : "HIDROCLOROTIAZIDA", "concentracion" : "25 mg" }, "total" : 780 }
{ "_id" : { "principio_activo" : "ATORVASTATINA", "concentracion" : "20mg" }, "total" : 690 }
{ "_id" : { "principio_activo" : "NAPROXENO", "concentracion" : "250 mg" }, "total" : 637 }
{ "_id" : { "principio_activo" : "ESOMEPRAZOL", "concentracion" : "20 mg" }, "total" : 630 }
MongoDB Enterprise drogueria-shard-0:PRIMARY>
```

Ilustración 17. Agregación desde la línea de comando

Fuente: Elaboración propia

Desde la línea de comandos es posible hacer la consulta anterior que se visualiza desde MongoDB Compass de una forma más técnica con la cual se puede comparar los resultados además de poder corroborar que estos fueron correctos y acorde con toda la información suministrada.

Agregación para los tops

```
([

    { $match :{ sexo:"F"}},

{$group: {_id:{ principio_activo: "$principio_activo", concentracion:"$concentracion"}, total:{$sum:
    "$cantidad"}}},

    {$sort: {total: -1}},

    {$limit: 10}

])
```

Conclusiones

La Droguería Puerto Boyacá S.A.S. es una mipyme que tiene como principal modelo de negocio la realización de contratos con entidades prestadoras de salud, para encargarse de la entrega de productos farmacéuticos y medicinales. Esta a pesar de llevar más de 10 años funcionando no ha logrado crecer, siendo además incapaz de determinar su ganancia o pérdida frente a esta actividad.

Al reunir y revisar los distintos archivos o conjuntos de datos suministrados por la droguería se evidenciaron problemas en la administración de los mismos, en donde los problemas principales que le impiden llegar a un nivel de madurez analítico descriptivo son; la falta o ausencia de datos y el uso de archivos pdf e imagénes para almacenar otros, lo que impide una realización de análisis completos y certeros que revelen información de valor.

Con el restante de archivos proporcionados, siendo todos estos de excel, se pudo especificar los conjuntos de datos pertinentes con los datos mínimos requeridos para llegar al nivel descriptivo en el que se pueda determinar todo lo que ocurre respecto al proceso en cuestión. Con estos conjuntos de datos se llegaron a análisis que caracterizan a la población atendida durante el mes de Febrero del presente año, y por otra parte, se llegó a un modelo de análisis comparativo entre la cantidad de productos entregados y los que como máximo podría llegar a entregar, determinando así los resultados de la ejecución del contrato únicamente para el mes de Febrero.

Finalmente se demuestran las ventajas y beneficios para una mipyme como la Droguería Puerto Boyacá del uso de MongoDB con en su versión gratuita de Atlas y el software open

source MongoDB Compass, para llegar de una forma rápida y sencilla a un estado descriptivo inicial en el cual puedan visualizar y consultar lo que ocurre. Para esto se muestran los análisis dados por el schema de Compass y se deja un modelo de querie que permita sacar los mismos análisis mostrados.

Se ha de destacar que los análisis mostrados son solo de lo ocurrido en el mes de Febrero, mientras que para un análisis más concluyente se debe presentar una completitud en los datos de los conjustos propuestos, esto ocurrió puesto que la Droguería Puerto Boyacá S.A.S no pudo suministrar la totalidad de los datos.

Con los conjuntos de datos y las herramientas mencionadas se puede llegar fácilmente a un nivel de madurez analítica descriptiva. Sin embargo, para que esto se de, se mantenga y evolucione, es necesario por parte de la empresa incluir personal especializado de TI o de los campos afines como ciencia de datos, big data y análisis de datos.

Referencias

Axelos. (2019). *ITIL® Foundation ITIL 4 Edition.*

Beth Chrissis, M., Konrad, M., & Shrum, S. (2009). *CMMI: Guía para la integración de procesos y mejora de productos.* Madrir: Addison Wesley.

Carvalho, J. V., Rocha, Á., Vasconcelos, J., & Abreu, A. (2018). A health data analytics maturity model for hospital information systems. *International Journal of Information Management*.

COLOMBIA FINTECH. (2019). *El 62% de las pymes colombianas no tienen acceso a financiamiento*. Retrieved from COLOMBIA FINTECH: https://www.colombiafintech.co/novedades/el-62-de-las-pymes-colombianas-no-tiene-acceso-a-financiamiento

Data Centric. (2017, Mayo 23). *Data Centric*. Retrieved from Data Centric. Business Customer Location Insight: https://www.datacentric.es/blog/geomarketing/data-visualization-analisis-datos/

Elliott, T. (2013). *GartnerBI: Analitics Moves To The Core*. Retrieved from Digital Business & Business Analytics: https://timoelliott.com/blog/2013/02/gartnerbi-emea-2013-part-1-analytics-moves-to-the-core.html

GAD. (n.d.). *MongoDB ¿qué es y cómo funciona?* Retrieved from SiliconNews: https://siliconnews.es/mongodb-como-funciona/

Gamma, C. O. (2012). Método para la madurez para la calidad de los datos. *Universidad Autónoma de Occidente*, 15.

Grupo Bit. (2018). *¿Cuál es el panorama de la analítica de negocios en Colombia?* Retrieved from Grupo Bit: https://business-intelligence.grupobit.net/blog/cual-es-el-panorama-de-la-analitica-de-negocios-en-colombia

Haan, L. D. (2018). The integration of Big Data in purchasing, as designed in new Big Data Purchasing Maturity model. *University of Tweente*, 64.

Lara, J. (2018, Diciembre 19). *Las 5 V del Big Data.* Retrieved from eadic: https://www.eadic.com/las-cinco-v-del-big-data/

López Briega, R. (2016). *Libro online de IAAR.* Retrieved from Libro online de IAAR: https://iaarbook.github.io/

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data. La revolución de los datos masivos.* (A. Iriarte, Trans.) Madrid: Turner Publicaciones S.L.

Ministerio de Salud. (2000, Diciembre 27). Resolución 3374 de 2000. *Resolución 3374 de 2000*. Bogotá, Colombia. Retrieved from https://docs.supersalud.gov.co/PortalWeb/Juridica/OtraNormativa/R3374000.pdf

MondoDB inc. (2008). *Agreggation*. Retrieved from MongoDB Manual: https://docs.mongodb.com/manual/aggregation/

MongDB. (n.d.). *MongoDB.* Retrieved from MongoDB: https://www.mongodb.com/es

MongoDB. (n.d.). *MongoDB Atlas*. Retrieved from MongoDB: https://www.mongodb.com/cloud/atlas

MongoDB. (n.d.). *MongoDB Compass*. Retrieved from MongoDB: https://www.mongodb.com/products/compass?lang=es-es

Muñoz, A. (2017, Noviembre 4). *¿Qué es Data Minig?* Retrieved from Computer Hoy: https://computerhoy.com/noticias/internet/que-es-data-mining-70663

Omedes, J. (2017, Febrero 1). *Analítica de aprendizaje, ¿cuál su nivel de madurez en ?* Retrieved from IAD Learning: https://www.iadlearning.com/es/analitica-de-aprendizaje-madurez/

pandas. (2011). *Python Data Analysis Library.* Retrieved from pandas: https://pandas.pydata.org/

Rodríguez Cruz, Y., & Pinto Molina, M. (2010). Evolución, particularidades y carácter informacional de la toma de decisiones organizacionales. *Acimed, 21(1),* 57-77. Retrieved from https://www.medigraphic.com/pdfs/acimed/aci-2010/aci101f.pdf

Rouse, M. (2012, Noviembre). *Análisis de datos.* Retrieved from SeachDatacenter: https://searchdatacenter.techtarget.com/es/definicion/Analisis-de-Datos

Saffirio, M. (2008, Junio 21). *Escala de Madurez - Prpcess Maturity Model*. Retrieved from Tecnologías de la Información y Procesos de Negocios (BPM): https://msaffirio.wordpress.com/2008/06/21/escala-de-madurez-%E2%80%93-process-maturity-model/

Sánchez Crespo, L., & Villafranca Alberca, D. (2000). Data Cleaning para Data Warehouse (tesis doctoral). Ciudad Real, Universidad de Castilla la Mancha. Retrieved from http://www.sanchezcrespo.org/DataCleaning.htm

Tanraparni, D., & Theodore, J. (2003). *Exploratory Data Mining and Data Cleaning.* Jhon Wiley & Sons.

Vidal, J. (2011, Julio 26). *Limpieza de datos*. Retrieved from DATAprix: https://www.dataprix.com/es/category/integracion-datos/integracion-datos/calidad-datos/limpieza-datos

(2019). Visión PYMES 2019. (B. I. Corporation, Interviewer) Retrieved from http://incp.org.co/Site/publicaciones/info/archivos/vision-pymes-2019-2552.pdf

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*(10).